Proceedings of the 2006 WSEAS International Conference on Mathematical Biology and Ecology, Miami, Florida, USA, January 18-20, 2006 (pp168-173)

Comparing RNA molecules based on their secondary structures

Na Liu [1,2*] Tianming Wang[2,3]

[1] *Department of Applied Mathematics , Dalian University of Technology, Dalian 116024, China*
[2] *College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, China*
[3] *Department of Mathematics , Hainan Normal University, Haikou 571158, China*

Abstract: In this paper, we present an alternative method to compare RNA molecules(not allowing for pseudo-knots). That means the basic bases and the base pairing are taken into account simultaneously, which is accomplished by using characteristic sequence of RNA molecule. Hence less information will be lost in the process of comparing. Moreover, the process is easy to operate and can give result rapidly. The validation is tested at the end of this paper.
Keywords: secondary structure, similarity, transition probability vector, characteristic sequence, hierarchical clustering

1. Introduction

Ribonucleic acids (RNAs) are important molecules and fulfill a broad range of functions. They have recently become the center of much attention. Comparing different RNA molecules or substructures of them is important work, which can provide valuable information for their structural prediction. Given an RNA molecule with structure known and a newly-discovered RNA molecule with structure unknown, we say that the newly-discovered RNA molecule may have the same structure as the former if the similarity degree between them is high. Or, given an RNA molecule with function known and an RNA molecule with function unknown, we may draw such a conclusion that the latter will probably have the same function as the former if their secondary structures are similar.

However, the formation of the hydrogen bonds between certain two bases makes it difficult to compare different RNA molecules efficiently. It challenges the researchers. Up to now, various methods have been proposed.

Usually, loops and stems are defined and considered in a tree model, where they are regarded as nodes and lines(or arcs). Obviously the information on the primary sequence—the order of bases is completely lost. Shapiro,B et al [1,2] prefer to construct tree model and have proposed several tree algorithms to compare the secondary structures of RNA molecules. The similar idea can be found in [3,4,5]. To overcome such shortage and avoid the lost of much information, several methods that consider primary sequence and the secondary structure have been developed, which can be found in[6, 7]. Hofacker et al[8] computed the base pair probability matrices of RNA molecules and then they compared those matrices.

In this paper, by means of linear sequences that contain the information on secondary structures of RNA molecules, we propose an alternative method to compare RNA molecules and therefore deduce the similarity relationship of those molecules. Here the transition probability vector is calculated based on characteristic sequence. It incorporates and contains the information on secondary structure and is a kind of numerical characterization of RNA molecule. Then the distance between RNA molecules is evaluated by using transition vectors. To test the validation of our method, we apply it to a set of data. The usage of the distance matrix constructed according to transition probability vectors to hierarchical clustering analysis is shown.

2. Methodology

*To whom correspondence should be addressed: Fax: +(86)411-84706100.
E-mail address: liunasophia@163.com, wangtm@dlut.edu.cn
The work is supported by the National Science Foundation of China(10571019)

2.1 Characteristic sequence of RNA molecule

As we know, the primary sequence of RNA molecule is a string over the alphabet {A, C, G, U }. Because of the formation of hydrogen bonds, certain two bases will pair with each other. As a consequence, RNA chain folds in space and form the secondary structure. To large extent, it is the secondary structure that determines the function of RNA molecule. So the information contained in secondary structure is important. At the same time, the bases(they have different chemistry property) and the order of them in the primary sequence affects the function of RNA molecule. Hence the information contained in primary sequence shouldn't be ignored.

The characteristic sequence is a linear sequence that characterizes an RNA molecule[9]. It indicates the order of basic bases and the status of base pairing simultaneously. In other words, it is constructed based on the secondary structure.

It is constructed as follows: Beginning from the 5'-terminal of RNA chain, we scan each base along the chain till its last base. If the base that is being scanned pairs with other base, then we substitute its upper case in bold for the base and the upper case for the other base in this base pair. Or else, the lower case is adopted. The same rule is followed when we go to scan the next base that hasn't been scanned. For example,for the structure as follows, which is represented by 'bracket notation',

CAGCAUCGCUCCUAAUACAA
..(((.....))).....((...)).....

its characteristic sequence is

c**AGC**aucGCUcc**UA**aUAcaa.

Obviously, the information that is displayed by 'bracket notation' has been shown by characteristic sequence. It is a concise and visual description of RNA molecule as viewed from its secondary structure. We denote the characteristic sequence by CS.

2.2 Transition probability vector

Given a biological sequence, what's the most critical is to abstract the numerical characteristics from it. The more information you abstract, the better your result will be. Here we use a vector to characterize an RNA molecule. Here two steps are needed:

Step1: Compute the transition probability based on a given CS. According to the construction of characteristic sequence, it is in fact a linear sequence defined over the alphabet { **A, C, G, U**, a, c, g, u, A, C, G, U }. Given a CS, we care the probability of the incident that a letter depends on its previous one. For a letter in the alphabet, it may appear in CS. Its following letter in CS may be **A, C, G, U**, a, c, g, u, A, C, G or U. For each case, we can calculate its probability. We call these probabilities transition probabilities. They are calculated by the following formula:

$$p_{a_i a_j} = \frac{n_{a_i a_j}}{\sum_{j=1}^{12} n_{a_i a_j}}, \qquad if \sum_{j=1}^{12} n_{a_i a_j} \neq 0 \quad i,j = 1,2,\cdots\cdots\cdots\cdots,12$$

$$p_{a_i a_j} = 0, \qquad if \sum_{j=1}^{12} n_{a_i a_j} = 0 \neq 0 \quad i,j = 1,2,\cdots\cdots\cdots\cdots,12$$

where a_i represents the ith element of alphabet { **A, C, G, U**, a, c, g, u, A, C, G, U }; $n_{a_i a_j}$ represents the occurrence times of the event that base a_i is followed by base a_j in CS.

Step2: Construct the transition probability vector TPV for RNA secondary structure. Given a CS, we arrange all the probabilities into a vector, denoted by TPV. The vector is constructed as follows:

$$TPV = (p_{a_1a_1} \;\; p_{a_1a_2} \;\; \cdots\cdots \;\; p_{a_1a_{12}} \;\; p_{a_2a_1} \;\; p_{a_2a_2} \;\; \cdots\cdots \;\; p_{a_2a_{12}} \;\; p_{a_{12}a_1} \;\; p_{a_{12}a_2} \;\; \cdots\cdots \;\; p_{a_{12}a_{12}})$$

TPV numerically describes an RNA molecule as viewed from its secondary structure. It may be regarded as an indicator. Take the structure listed above for example, its TPV is (0 0 0.5 0 0.5 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.3 0 0 0.3 0 0 0 0.3 0.2 0 0 0.2 0.2 0.2 0 0 0 0 0.2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0.5 0 0 0.5 0 0 0)

2.3 Evaluating the difference of different RNA molecules

In essence, each RNA molecule is represented by a vector. Then the distance between any two RNA molecules can thus be measured by the Euclidean distance of the two vectors. Now the comparison of RNA molecules becomes the comparison of vectors, which greatly decreases the complication that exists in many methods.

Given two RNA molecules A and B, whose transition probability vectors are denoted by TPV_A, TPV_B, the Euclidean distance of them is:

$$E_{AB} = \sqrt{\sum_{k=1}^{144} (V_{Ak} - V_{Bk})^2}$$

where V_{Ak} represents the kth component of TPV_A. E_{AB} measures the distance between A and B; Obviously the smaller E_{AB} is, the more similar A and B are.

3. Results and discussion

From what we have introduced above, given n RNA molecules $S_1, S_2, \cdots\cdots, S_n$, one may obtain their $TPVs$ easily. Then the distance between any two RNA molecules, $E_{ij}, i, j - 1, 2, \cdots\cdots, n$, can be obtained based on $TPVs$.

To test the validation of our method, we analyze a set of RNA molecules by our method. The data set is composed of nine 5S RNA molecules from Eukaryotes and Archaebacteria. They are Halobacterium spl(Hal-s), Pyrodictium occultum(Pyr-o), Sulfolobus spl(Sul-s), Actinia equina(Act-e), Basidiobolus magnus(Bas-m), Diatoma tenue(Dia-t), Chrysaora quinque(Chr-q), Christiansenis pallida(Chr-p) and Planocera recticulata(Pla-r).

After computing the pair distance E_{ij}, we arrange all the values into a matrix for clear and systematical display. This matrix is listed in the form of Table as follows, Table1. One may obtain more information about their similarity from the matrix.

Table The distance matrix derived from transition probability vectors

Names	Hal-s	Pyr-o	Sul-s	Act-e	Bas-m	Dia-t	Chr-q	Chr-p	Pla-r
Hal-s	0	1.816	1.321	1.754	1.531	1.573	1.787	1.600	1.470
Pyr-o		0	1.481	1.678	2.021	1.819	1.861	1.762	1.792
Sul-s			0	1.468	1.522	1.347	1.463	1.433	1.235
Act-e				0	1.266	1.117	0.896	1.109	0.852
Bas-m					0	0.973	1.252	0.957	1.160
Dia-t						0	0.98153	1.080	0.967
Chr-q							0	1.257	0.815
Chr-p								0	1.047
Pla-r									0

Another usage of such matrix is that it may be used to hierarchical clustering analysis. The quality of a clustering analysis may show if the matrix is good and therefore if our method of abstracting information from RNA molecules is efficient.

The result of the hierarchical clustering analysis of these nine RNA molecules by our method is shown in Figure1. Clearly, the molecules from the Eukaryotes are grouped into one cluster and the molecules from the Archaebacteria are grouped into the other cluster. In each cluster, the similarity relationship is further shown. we observe that: 1. Actinia equina,Chrysaora quinque and Planocera recticulata are grouped closely (they belong to Animalia); 2. Basidiobolus magnus and Christiansenis pallida are grouped closely (they belong to fungi); 3. Pyrodictium occultum , Sulfolobus spl and Halobacterium spl are grouped closely (they belong to Archaebacteria). This is consistent with the results got by others[10,11].

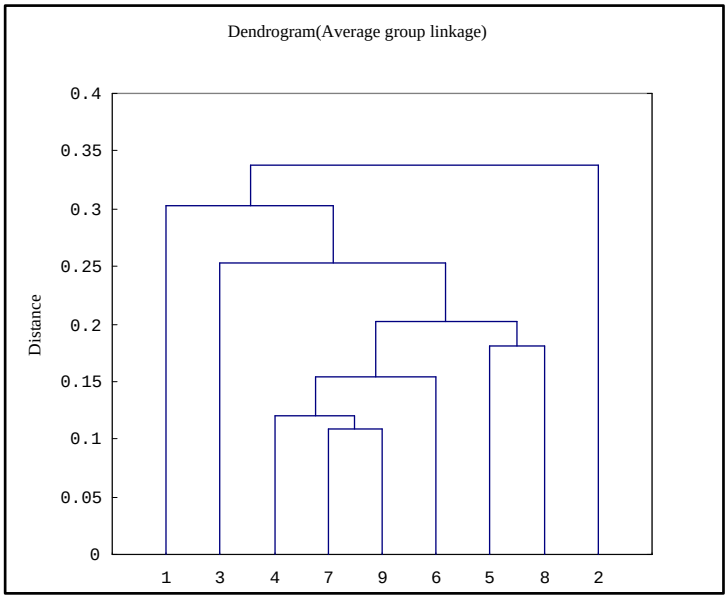

Figure1. Dendrogram of the hierarchical clustering of 12 secondary structures. 1-Halobacterium spl, 2-Pyrodictium occultum, 3-Sulfolobus spl, 4-Actinia equina, 5-Basidiobolus magnus, 6-Diatoma tenue, 7-Chrysaora quinque, 8-Christiansenis pallida, 9-Planocera recticulata

Note that our method can be used in not only complete molecules, but also the substructures of RNA molecules when we focus on their local structures. Here we add four 3'-terminal substructures into the above data set. They are the 3'-terminal of alfalfa mosaic virus(AlMV-3),citrus leaf rugose virus(CiLRV-3),lilacring mottle virus(LRMV-3),prune dwarf ilarvirus(PDV-3). Their structures are shown in Figure2.

By using our method, we analyze the relationship of these thirteen structures. Figure4 shows the result of their hierarchical clustering analysis, which is derived from their transition probability vectors. Referring to Figure2, we see the method is robust. And the result is reasonable.

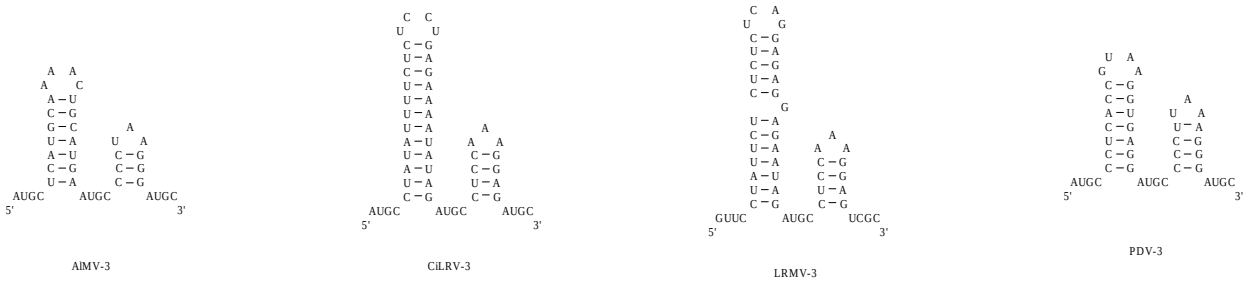

Figure2. The 3'-terminal structures of four viruses.

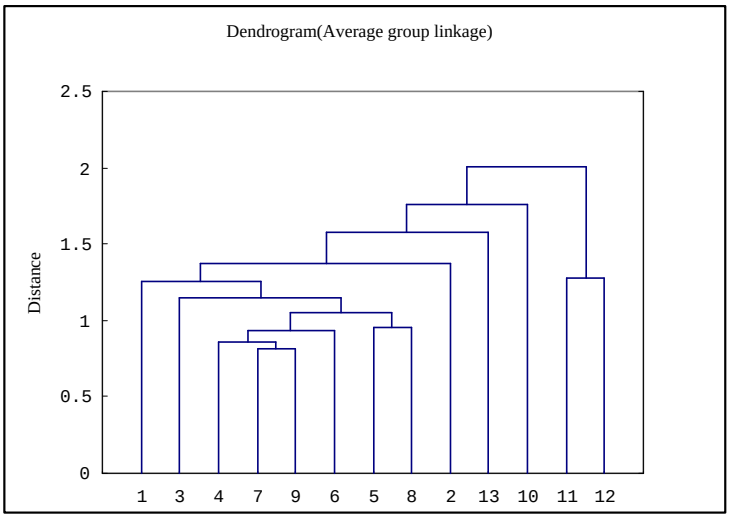

Figure3. Dendrogram of the hierarchical clustering of thirteen secondary structures. 1-Halobacterium spl, 2-Pyrodictium occultum, 3-Sulfolobus spl, 4-Actinia equina, 5-Basidiobolus magnus, 6-Diatoma tenue, 7-Chrysaora quinque, 8-Christiansenis pallida, 9-Planocera recticulata, 10-AlMV-3,11-CiLRV-3,12-LRMV,13-PDV-3

4. Conclusions

With more and more sequences and structures available, powerful computational methods are needed to analyze them. The similarity analysis can help determine the structure of sequence or the function of structure. Here we deal with RNA molecules(not allowing for pseudoknots). In this paper, we propose a new method to compare RNA molecules in terms of secondary structures. It makes use of transition probability vectors that characterize RNA molecules. Therefore the comparison of RNA molecules is transformed into the comparison of vectors. From the result of its application to 5S RNA molecules, we say that it can obtain reasonable result, *i.e.* our method is feasible for comparing RNA molecules and deduce their similarity relationship. Furthermore, the whole process is easy to operate and it can give the result rapidly. Note that the lengths of compared molecules are not restricted.

ACKNOWLEDGEMENTS

The authors thank Maciej Szymanski sincerely for his providing us with the secondary structures of all these 5S rRNAs.

References

[1] Shapiro,B. An algorithm for comparing multiple RNA secondary structures. *Comput. Appl. Biosci.*, **4** (3),(1988), 387-393.

[2] Shapiro,B.,Zhang,K. Comparing multiple RNA secondary structures using tree comparisons. *Comput. Appl. Biosci.*, **6** (4),(1990) , 309-318.

[3] Le,S.Y.,Nussinov,R.,Maizel,J.V. Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res.*,**22**, (1989), 461-473.

[4] Le,S.Y.,Owens,J.,Nussinov,R.,Chen,J.H.,Shapiro,B.,Maizel,J.V. RNA secondary structures: comparisons and determination of frequently recurring substructures by consensus. *Comput. Appl. Biosci.*, **5**,(1989), 205-210.

[5] Dulucq,S.,Tichit,L. RNA Secondary structure comparison: exact analysis of the Zhang–Shasha tree edit algorithm. *Theoretical Computer Science.* **306**,(2003), 471 - 484 .

[6] Bafna,B.,Muthukrishnan,S.,Ravi,R. Comparing similarity between RNA strings. *Proc. Combinatorial Pattern Matching Conference 95, Lecture Notes in Computer Science* . **937**,(1995), 1-14.

[7] Corpet,F.,Michot,B. RNAlign program: alignment of RNA sequences using both primary and secondary structures. *Comput. Appl. Biosci.* **10**,(1995), 389-399.

[8] Hofacker,I.L.,Bernhart,S.H.F.,Stadler,P.F. Alignment of RNA base pairing probability matrices. *Bioinformatics.* **20**,(2004), 2222-2227.

[9] Liu,N.,Wang,T.M. An alignment-free approach for comparing RNA secondary structures without pseudoknots.(2005), Submitted.

[10] Hiro,H.,Osawa,S. Evolutionary change in 5S rRNA secondary structure and a phylogenetic tree of 54 5S rRNA species. *Proc. Natl. Acad. Sci USA.* **76** ,(1979),381-385.

[11] Hiro,H.,Osawa,S. Evolutionary change in 5S rRNA secondary structure and a phylogenetic tree of 352 5S rRNA species. *Proc. Natl. Acad. Sci USA.* **19** , (1986), 163-172.